Express Mailing Label No. EL487542238US

PATENT APPLICATION
Docket No. 1785.2.2

UNITED STATES PATENT APPLICATION

of

David C. Taylor

for

USER-CONTEXT ANALYSIS ENGINE

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

USER-CONTEXT ANALYSIS ENGINE

BACKGROUND OF THE INVENTION

1. Related U.S. Applications

This application claims priority to U.S. Provisional Patent Application Serial No. 60/146,878, filed August 3, 1999, entitled "Data Extraction Tool," which application is incorporated herein by reference.

2. The Field of the Invention

This invention relates to a data extraction tool and, more particularly, to novel systems and methods for organizing information from a database for ready access by a user.

3. The Relevant Technology

In what is known as the information age, information is readily available electronically, through information repositories known as datastores and databases. Datastores are substantially unorganized collections of data, while databases are indexed in some fashion. The Internet, the world's largest database, has made available enormous quantities of information to anyone with a personal computer and Internet access. This can be very helpful for people who wish to learn about something or conduct business in the convenience of their own homes. However, it can also be tremendously time-consuming to locate a desired bundle of information among the millions available.

The Internet is organized only by the name of each web site. Each individual or group maintaining a web site decides how that web site will be organized. Thus, there is no official catalog of information available on the Internet. Anyone desiring information must hypothesize which web sites would be likely to have the desired data and navigate through

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

those web sites according to the organization set up by the web site's operator. Although other databases and datastores are small, many exhibit the same organizational difficulties.

Some companies have developed portals to automate a portion of the search for information. Most of these portals are text-based. Currently available portals include search engines, and directories.

To use a search engine, a user provides a set of words to search for, and the search engine returns a list of "hits," or web sites containing those words. Search engines are advantageous in that they require little user input or understanding of the operation of the search engine. However, they can be difficult to work with for a number of reasons.

For example, the list may contain a vast number of hits, few of which actually relate to the desired piece of data. Conventional keyword searching returns any instance of the word being sought, regardless of the way the word is used in the web site. Although a user may add additional keywords to narrow the search, there often is no combination of words that must be found together to exclude all irrelevant pages while keeping all relevant ones.

Also, many conventional search engines return only the home page of a web site that contains the keyword. It is then up to the user to find the keyword in a site and determine whether it is relevant. This requires a user to figure out how the site is organized and follow the right links. This can be difficult because there may be no links that clearly indicate where the keyword is.

The output from most search engines is simply a page of links to possibly relevant sites. A user may wish to supplement or rearrange the search results, but the way the results of a search are formatted typically makes addition or modification of criteria difficult or impossible.

Moreover, information obtained through a search often becomes outdated. Currently, a user must revisit previously found sites to determine whether the old information

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

is still valid. Additionally, a user must perform a new search to locate any newly relevant sites and search through those sites for relevant information.

Directories function differently than search engines. Rather than search based on keywords provided by a user, most directories provide a user with an information scheme, often hierarchically organized. The user then chooses what type of information to search for, designating narrower groups of information with each choice. Ultimately, the user reaches the bottom level of the hierarchy and receives a list of links to information within that level.

Directories are advantageous in that information concerning a certain topic is typically grouped together. A directory probably will not inundate a user with information, but rather provide a few links believed to be important by the creators of the directory. Nevertheless, directories have drawbacks of their own.

For example, traditional directories contain information deemed of value by those who compile them. A user may have an entirely different view of what is important and what is irrelevant. A user may thus find that information he or she needs simply is not available on the directory.

Also, directories take time to navigate. A user must make a series of decisions to reach any useful information at all. Even then, a user may find it necessary to backtrack and choose a different route through the hierarchy. Since a user cannot fashion groupings of information, he or she may be required to view several branches of the hierarchy to obtain the full range of information he or she desires.

Moreover, if a user does not know how to classify the bit of information sought, he or she may not even be able to find it in the directory. For example, a user desiring to find the meaning of "salmonella" in a biological directory may spend great amounts of time looking through the "aquatic life" branch of the directory, without ever realizing that "salmonella" is more properly classified as "microscopic life." The more a user's view of how

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

information should be organized differs form that of the directory's creators, the more difficult it will be for the user to find information in the directory.

Consequently, there is a need for a data extraction tool capable of providing many of the benefits of both search engines and directories, without drawbacks listed above. For example, there is a need for a tool that could reliably provide a list of highly relevant information locations based on a simple text query. Furthermore, such a tool should provide ready access to the exact location of the information. Preferably, the tool would supply the user with a list of locations or links that can be easily sorted and updated for the convenience of the user. Furthermore, the tool should not require that the user understand the configuration of the tool's internal databases.

In addition to the problems mentioned above, current searching methods are deficient in a number of other ways. Consequently, a more advanced data extraction tool may provide numerous benefits to those desiring to obtain information from a large datastore or database, such as the Internet.

OBJECTS AND BRIEF SUMMARY OF THE INVENTION

In view of the foregoing, it is a primary object of the present invention to provide a data extraction tool capable of context-sensitive searching, pinpointing, databasing, automatically updating information, or any combination thereof, from a datastore or database. Consistent with the foregoing object, and in accordance with the invention as embodied and broadly described herein, an apparatus and method are disclosed, in suitable detail to enable one of ordinary skill in the art to make and use the invention. In certain embodiments, an apparatus and method in accordance with the present invention may be directed to locating prices for products that can be purchased over the Internet.

Selected embodiments of the invention have modules for mining, databasing, searching, filtering, pinpointing, presenting, indexing, or updating information for a user, or

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

some combination thereof. According to some embodiments, the data extraction tool mines information from the information source and organizes the locations of that information within a local database. Then, a user may query the tool for a desired type of information. The tool filters the local database to provide a set of pinpoint site locations with information of the type requested in the query. These pinpoint site locations are presented to a user and indexed for future reference. The index of site locations may be updated automatically by the tool.

A context system is provided for manually or automatically determining the proper context for a user's query. Thus, the data extraction tool provides information with a high probability of relevance to the user. The user obtains the information without expending much effort to refine the search.

These and other objects, features, and advantages of the present invention will become more fully apparent from the following description and appended claims, or may be learned by the practice of the invention as set forth hereinafter.

BRIEF DESCRIPTION OF THE DRAWINGS

In order that the manner in which the above-recited and other advantages and objects of the invention are obtained will be readily understood, a more particular description of the invention briefly described above will be rendered by reference to specific embodiments thereof which are illustrated in the appended drawings. Understanding that these drawings depict only typical embodiments of methods and apparatus in accordance with invention and are not therefore to be considered to be limiting of its scope, the invention will be described and explained with additional specificity and detail through the use of the accompanying drawings in which:

Figure 1 is a schematic block diagram of an individual node in a network system;

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

Figure 2 is a schematic block diagram of one embodiment of a data extraction tool on a server, such as the server of Figure 1;

Figure 3 is a schematic block diagram of a data extraction tool, as shown in Figure 2, configured for use with a node in a network system, as shown in Figure 1;

Figure 4 is a schematic block diagram of data structures for administering and executing a user interface in accordance with the invention;

Figure 5 is a schematic block diagram of data structures for administering and executing a filtering module in accordance with the invention;

Figure 6 is a schematic block diagram of data structures for administering and executing an attributes index in accordance with the invention;

Figure 7 is a schematic block diagram of methods for implementing one embodiment of the data structures and functions of Figure 2 in accordance with the invention;

Figure 8 is a schematic block diagram of methods for implementing one embodiment of the mining step of Figure 7 in accordance with the invention;

Figure 9 is a schematic block diagram of methods for implementing one embodiment of the database construction step of Figure 7 in accordance with the invention;

Figure 10 is a schematic block diagram of methods for implementing one embodiment of the searching step of Figure 2 in accordance with the invention;

Figure 11 is a schematic block diagram of an alternative method for implementing the data structures and functions of Figure 2;

Figure 12 is a schematic block diagram of methods for implementing one embodiment of the context construction module of Figure 11 in accordance with the invention;

Figure 13 is a schematic block diagram of methods for implementing one embodiment of the context comparison module of Figure 11 in accordance with the invention;

Figure 14 is a schematic block diagram of methods for implementing one embodiment of the information matching module of Figure 11 in accordance with the invention; and

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

Figure 15 is a schematic diagram of a hierarchical database usable in conjunction with the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The presently preferred embodiments of the present invention will be best understood by reference to the drawings, wherein like parts are designated by like numerals throughout. It will be readily understood that the components of the present invention, as generally described and illustrated in the figures herein, could be arranged and designed in a wide variety of different configurations. Thus, the following more detailed description of the embodiments of the apparatus, system, and method of the present invention, as represented in figures 1 through 15, is not intended to limit the scope of the invention, as claimed, but is merely representative of presently preferred embodiments of the invention.

Those of ordinary skill in the art will, of course, appreciate that various modifications to the details of the figures may easily be made without departing from the essential characteristics of the invention. Thus, the following description of the figures is intended only by way of example, and simply illustrates certain presently preferred embodiments consistent with the invention as claimed.

Referring now to Figure 1, a system 10 or network 10, such as the Internet, may include nodes 11 (e.g. nodes 50, 52, 54). Each node 11 may include a processor 12 and memory devices 14, such as storage devices 16, read only memory (ROM) 18, and random access memory (RAM) 20, sometimes referred to as operational memory. The node 11 may include a variety of input devices 22, and output devices 24 whether dedicated as illustrated in Figure 1, or more generally available over a network.

Typically, a node 11 may include a network card 26 for connecting to a network 30 (e.g. network 10) outwardly, and a bus 32 for interconnecting elements internally.

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

Input devices 22 may include a keyboard 34, a mouse 36 or other pointing device 36 such as a stylus or graphics tablet, an interactive touch screen 38, a scanner 40, or even a storage device 41 for providing data to the node 11. Similarly, output devices 24 may include monitor 42, printer 44, storage devices 46, and the like for providing data from the node 11.

A router 48 may interconnect networks 30, 50 where each network 30, 50 may include some simple nodes 52, such as clients 52a-52d, and servers 54. Networks 30, 50 are well understood in the art. Accordingly, the hardware illustrated is by way of example, and not limitation as to the hardware suite on which the invention may be implemented. More or less equipment may be used in many particular embodiments.

The system 10 is the datastore or database from which information is to be obtained. However, the system 10 need not be configured as shown in Figure 1. For example, the system 10 may be a database contained on a single computer. However, many of the subsequent descriptions will refer to the system 10 as a distributed network 10 of computers, such as the Internet.

Figure 2 shows one embodiment of a data extraction tool 110, or tool 110, with its associated modules. A mining module 112 gathers information from a data source, preferably the Internet. A databasing module 114 categorizes and sorts information within a local database. This information can be actual data directly from the data source, or it can be simply pointers to locations of data within the data source.

An input module 116 interfaces with a user to receive a request for information. A filtering module 118 filters information to isolate the data most relevant to a user's request. A pinpointing module 120 locates and returns identification of the exact location of information. A presentation module 122 presents information summaries and locations to a user. An indexing module 124 organizes information for use and access by a user. An updating module 126 automatically updates information in a local database.

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

The arrows in Figure 2 show a general chronological flow. However, the modules shown do not have to be accessed in the order shown. In addition, modules can operate multiple times. For example, the updating module 126 may utilize the mining module 112 to obtain more information, which may then be shown to a user by the presentation module 122 and stored by the indexing module 124.

Referring to Figure 3, a node 11 may have a hard disk 128 or HD 128, an input/output port 130 or I/O 130, a central processing unit 132 or CPU 132, and a memory 133. The modules 112, 114, 116, 118, 120, 122, 124, and 126 may be temporarily stored for use in the memory 133, permanently stored in the hard disk 128, and processed through the central processing unit 132. A user interface 134 and a server link 136 provide for communication with a user and with the network 10 via the I/O 130. A transaction interface 138 may also be included to permit purchasing and selling over the network 10.

Figure 4 shows some data structures that may be included in the user interface 134. A home page 140 provides a user with access to the node 11 with the data extraction tool 110. A web server 142 stores data that makes up the home page 140. CGI scripts 144 may display data from the web server 142 for a user in the form of text 146 and graphics 148.

A free form input module 150 receives searching parameters, in the form of a query, from a user. A semantic analysis module 152 parses the query and uses context templates 154 to develop a list of contexts that may correlate to information desired by a user. An inquiry module 156 prompts a user for further guidance concerning which of the available contexts are truly relevant. A query modification module 158 modifies the query to suit a user's response to prompting from the inquiry module 156. A presentation module 160 displays search results in a clear fashion for a user, and can also permit a user to interact with other nodes (e.g. nodes 50, 52, 54) where further information may be stored.

In addition, a site interaction module 162 can permit partial processing of information by the data extraction tool 110 before presentation to a user. A pinpoint selection module

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

164 chooses relevant sites for further processing. A login module 166, if needed, may permit the site interaction module 162 to automatically log onto a site where relevant information is stored. A link selection module 168 chooses the most relevant path within the site for retrieval of the desired information. A page parsing module 170 determines whether text from the site is relevant to a user's query.

Referring to Figure 5, some data structures that may be used in the filtering module 118 are shown. The semantic net 174 is a resource for matching query text from a user to text from a web site. Context clues 176 provide information for contextual comparisons based on classifications 178 of contexts in which a word may be found. A context selector 180 selects those contexts that correlate to the proper context for the query and isolates them via filters 182. The filters 182 may reference the context system 184, which simply provides a list of actions corresponding to each instance of a word. For example, the context system may specify that a site should be retained if a keyword is found in a certain context within the site, but that the site should be filtered out if the keyword is used in a different context.

Referring to Figure 6, the context clues 176 and the context selector 180 may reference an attributes index 185. The attributes index 185 contains a list of words 186. Each word has at least one meaning 187 indexed to that word, and each meaning 187 has a list of relations 188, such as synonyms, antonyms, subsets, supersets, usage correlation, and usage association. A second meaning 189, and however many meanings exist for the word 186, may also be included with an associated list of relations.

The relations 188 provide context clues 176 so that a given web site can be classified by context. The context may be determined, for example, by the frequency and combination of relations 188 that appear within the web site. Thus, the filters 182 can filter out those web sites in which the proper keyword is used in an irrelevant context.

Figure 7 shows one method for implementation of the data structures of Figure 2. In a mining step 190, a data extraction tool 110 mines information from the network 10. The

locations of the information, and possibly some of the information itself, may be compiled and formatted in a database construction step 191. A searching step 192 permits a user to query for information stored by the database construction step 191. A filtering step 193 selects the information most relevant to a user's query. A pinpointing step 194 determines the exact location of the relevant information on the network 10. A presenting step 196 organizes relevant information and provides it to a user. An indexing step 198 links relevant information to the location of that information on the network 10. An updating step 200 subsequently performs an automatic search of the network10 for new information relevant to the user's query.

Figure 8 shows possible steps that might be included within the mining step 190. In a route selection step 202, the tool 110 chooses an orderly method for processing information from the network 10. Preferably, the route selection step 202 involves an orderly progression to ensure that each potentially relevant parcel of data is processed once and only once. In an autonavigation step 204, the tool 110 receives information from the network 10 for processing in a content reading step 206.

In an evaluation step 208, the tool 110 evaluates the potential relevance of the text 146 of a site to future queries of a user. The tool 110 may be directed towards acquiring a certain type of information, or broadly used to obtain and categorize a wide variety of data. The scope of data to be mined determines how selective the evaluation step 110 will be. In a content extraction step 210, potentially relevant content is compared against a listing of needed information to further filter it in a database filtration step 212. The data are indexed for ready access by an addition to a master index step 214.

Figure 9 shows possible steps that might be included within the database construction step 191. A database structuring step 216 provides the structure and organization for the information. In a schema provision step 218, a relations recording step 220 and an indices recording step 222 organize data into fields that are appropriately linked together and indexed

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

for rapid reference. In an input data step 224, the tool 110 receives information gathered during the mining step 186.

A data classification step 226 uses discrimination functions 228 to categorize information within the schema developed by the schema provision step 218. A schema refining step 230 permits revision of the schema as needed to accommodate information that otherwise cannot be appropriately categorized within the schema. In a records filling step 232, the tool 110 adds data to form records.

In an addition to database step 234, the tool 110 adds data retrieved by the mining step 186 to the local database. This involves a number of steps. In a site identification step 236, the tool 110 identifies sites of relevant information. In a site isolation step 238, the tool 110 further filters sites based on criteria provided by a user or by the programming of the tool 110. For example, the tool 110 can be programmed to isolate sites capable of conducting commerce over the Internet. In that case, the site isolation step 238 would filter out all sites without a method for conducting commerce through the site. In a site contents classification step 240, the tool 110 classifies data into appropriate categories, as laid out in the schema. A data selection step 242 chooses classifiable data for transmission to a record preparation step 244, where data is added to records in the local database.

Referring to Figure 10, a number of steps may be included within the searching step 192. A user may request information by entering free form text or other query inputs in a query receiving step 246. In a query parsing step 248, the query is compared against a list of possible contexts by a semantic net reference step 250. In an inquiry preparation step 252, the tool 110 forms a question for a user, in a question selection step 254, to ask for clarification concerning which of the potential contexts that may match the query is the most relevant.

The inquiry computation step 256 may provide an estimate of the time required to perform a search for each potential context, so that a user will know how long the tool 110

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

will take to process a search for a given context. This is especially helpful when a user has provided a very broad query. In such a case, the computation time will be high, so a user will know that the search will take a comparatively long time and provide a comparatively large amount of information, perhaps more than desired.

In an additional input receiving step 258, the tool 110 receives more text or menu selections from a user to identify which of the context or contexts are desired for searching. In an index reading step 260, the tool 110 reads an index of information contexts created in conjunction with the database construction step 191. The relevant context or contexts in the index are linked to site locations for information. The tool 110 returns these site locations to a user in a pinpointed sites returning step 262.

Referring to Figure 11, the searching and filtering modules may alternatively be embodied as shown in Figures 11 through 14, in contrast to the configurations shown in Figures 5, 6, and 10. As above, the input module 116 may transmit text 117 reflecting a search query to the filtering module 118, which may then filter information to isolate what a user is seeking. In this embodiment, the filtering module 118 includes a context construction module 300 for assembling micro-contexts 301 based on the text 117, a context comparison module 302 for converting the micro-contexts 301 to macro-contexts 303, and an information matching modules 304 for matching the macro-contexts 303 to specific information 306 responsive to the user's query. The presentation module 122 again provides the information to a user.

The input module 116 may acquire text to describe information sought by a user in a variety of different ways. For example, a simple free form text search may be used, wherein the user types a query in plain language. Alternatively, a user may provide key words separated by operators such as and, or, not, and others known in the art. The input module 116 may be configured to refine the text through questions to be answered by a user. The filtering module 118 then receives the text from the input module 116. Until processed, the

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

text is only a series of words with no inherent meaning to a computer. The filtering module 118, in this embodiment, will convert the text into searchable portions to find matching information of the type desired by a user.

Referring to Figure 12, the context construction module 300 is shown in greater detail. The context construction module 300 assembles the words to form small, coherent groups, or micro-contexts 301, for examples, they may contain about 1 to 5 words. This is accomplished in part by using a block parser 316, which breaks down and interprets the text. The text can be broken up by the block parser 300 in a number of different ways. Key words 312 and their modifiers, if designated by a user, can form or define natural contexts for searching. Similarly, relative values 314 or priorities assigned to words in the text may be used by the block parser 310 to create micro-contexts 301. Occurrence patterns 316 may be used to form natural separations between groups of words.

These occurrence patterns 316 may be obtained from a user's history 318 corresponding to a given user's activities with the tool, including prior searches and results, or from a general language database such as the attributes index 185. The user history 316 in any case may provide the tool 110 with information concerning what information a user has requested in the past, and therefore what information the user is most likely looking for with a new inquiry.

Referring to Figure 13, the context comparison module 302 is shown in greater detail. The context comparison module 302 receives the micro-contexts 301 from the context construction module 300 and compares them to a corpus 330 of information. The corpus 330 may simply be a database with samples of information 332 in natural language format, indexed according to macro-contexts 303. These macro-contexts 303 may be more specific than the micro-contexts 301.

The corpus 330 is sized to suit the amount and type of information on the network 10. The corpus 330, for example, may be composed of portions of text from 100,000 to

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

200,000 web sites, or more, with each portion matched to a macro-context 303. The entire corpus 330 may be between 10 Megabytes and 10 Gigabytes in size, or larger. A text comparison algorithm 336 may be provided to match text from the corpus 330 to the micro-contexts 301, and then return the corresponding macro-context 303. The text comparison algorithm 336 may combine several micro-contexts 301 to permit a more specific search, thereby narrowing the number of matching macro-contexts 303.

Ideally, the context comparison module 302 will return a small number of macro-contexts 303. However, this may not be possible for two reasons. First, if multiple, important, micro-contexts 301 are identified by the context construction module 300, they might not appear together within any portion of the corpus 330. In such a case, the context comparison module 302 may return a series of macro-contexts 303 that match some fraction of the important micro-contexts 301. Although these macro-contexts may not precisely match a user's query, they may be ranked in order of likelihood that they will be relevant to the user. The ranking may be obtained by using the user history 318 and other factors, such as the number, probability, or nature of prior requests of the macro-context 303 by other users, to determine the probability that a given macro-context 303 is relevant to the user.

Alternatively, the micro-contexts 301 may not even be found in the corpus 330. In that case, a user may be referred to a user tracking module 338, which provides a user with portals to access and search the network 10 directly. The user tracking module 338 permits the tool 110 to track a user's progress through the network 10 to obtain further context information for the current search, acquire more general information regarding contexts important to the user, or find important information not currently present within the corpus 330.

A rapid mining module 339 may also be accessed while the user tracking module 338 is operating, to add nodes 52, or sites 52, to the corpus 330 and to process them through the

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

databasing module 114 "on the fly," or while the user is accessing the tool 110. These may be sites 52 visited by the user or suggested by the user's query.

After searching the network 10 through the user tracking module 338, a user may once again be referred to the input module 116, in order to provide additional text inputs, or the context comparison module 302 may resume operation to process the micro-contexts 301 through new additions to the corpus 330.

Referring to Figure 14, the information matching module 304 receives macro-contexts 303 from the context comparison module 302 and compares them to an indexed database 350. The indexed database 350 contains specific information 306 of the type desired by the user, indexed by macro-contexts 354 identical or similar to those provided by the context comparison module 302. Thus, using a structure-matching algorithm 356, the information matching module 304 can find the portion of specific information 306 that correlates to the macro-contexts 303 provided by the context comparison module 302. The specific information 306 located by the information matching module 304 may then be returned to the presentation module 122 to be presented to a user.

The presentation module 122 is preferably flexible in its operation. For example, the depth and breadth of specific information 306 returned may be varied according to a user's preferences. Once the specific information 306 is located within the indexed database 350, proximate information is easily gathered and returned. The order and arrangement of specific information 306 displayed may also be determined manually by a user or automatically by reference to the user history 318.

Referring to Figure 15, the indexed database 350 may be structured as a hierarchical database 400. The hierarchical database 400 is configured like a tree, with general information at the top and more specific information below. A parcel of information 402 desired by a user is a specific portion, and is therefore near the bottom of the hierarchical database 400. According to traditional methods prior to the current invention, a user would

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

locate the parcel of information 402 by navigating through the broadest classification 404 and through the branches 406, 408, and 410. A user might find this path difficult or even impossible to find, particularly if the user knows little about the parcel of information 402, the organizational scheme in which it resides, or the related elements in the hierarchy, and therefore little about where it should be classified.

The current invention permits a user to navigate across hierarchies straight to the parcel of information 402. The hierarchical database 400 remains transparent to the user, who need not familiarize himself or herself with the structure of the hierarchical database 400. Thus, the method disclosed herein provides horizontal navigation across a hierarchical database, in which the tool 110 intelligently determines exactly what the user is looking for and searches among the more specific, lower branches of the hierarchical database 400 to find it.

One application of such a hierarchical, searchable database is to provide information about products for sale over the Internet. In such a case, the presentation module 122 ultimately returns words to the user to denote the various products in the hierarchical database 400 that match the user's request. The presentation module 122 may, for example, be configured to sort products matching the user's request by brand, model, specifications, price, vendor, availability, distance to the vendor from the user, shipping cost, or any number of other relevant parameters.

In addition, the login module 166 may operate to navigate a site 52 for a user, including forms presented by the site 52 to collect information from the user. Thus, not only is a user freed from the need to navigate the hierarchical database, the user may also be permitted to access the site 52 and conduct business on it without having to navigate the structure of the site 52.

The tool 110 as configured above is also well adapted for use without such a hierarchical structure. The context matching capabilities of the tool 110 make the tool 110

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101

effective for improving the relevance and completeness of results to a query, regardless of what formats are used by the tool 110 to maintain and organize a local database. This is a vast improvement over current search engines, which typically search only for the exact text provided by the user, and thus deliver results that include irrelevant items and fail to include important information.

From the above discussion, it will be appreciated that the present invention provides a data extraction tool for extracting information from an information source. Extracted information is cataloged and indexed for future searching by a user. Although not limited to commerce, the method disclosed herein may be adapted to search for commerce-ready web sites on the Internet.

The present invention may be embodied in other specific forms without departing from its structures, methods, or other essential characteristics as broadly described herein and claimed hereinafter. The described embodiments are to be considered in all respects only as illustrative, and not restrictive. The scope of the invention is, therefore, indicated by the appended claims, rather than by the foregoing description. All changes that come within the meaning and range of equivalency of the claims are to be embraced within their scope.

What is claimed and desired to be secured by United States Letters Patent is:

MADSON & METCALF, P.C.
ATTORNEYS AT LAW
900 GATEWAY TOWER WEST
15 WEST SOUTH TEMPLE
SALT LAKE CITY, UTAH 84101